1 Reviewer 1

Dear Reviewer 1,

Thank you for this new review. The manuscript has been extensively revised to take into account the comments of Reviewer 3 and we hope that it still meets your original requirements. In what follows, we address your remaining questions and comments. To be noted that some of them refer to parts of the manuscript that have been removed in this new version. Minor edits have been directly incorporated to the manuscript.

I have a remaining comment regarding the input samples. The authors created an initial LHS sample of size 4000 to screen the influential parameter. Then they ranked these influential parameters based on a new sample of size 1000. Why is it required to create a new sample for the ranking? Could not the initial sample of size 4000 be used for this purpose (only considering the influential input and dropping the dimensions corresponding the non-influential inputs)? I think this should be clarified in the manuscript, since creating a new sample largely increases the computational cost of the analyses.
Indeed, this strategy could be applied to limit the computational budget, especially since dropping the dimensions of some inputs would not affect the LHS structure of the sample. In this study, as the scenario was of limited size we could afford additional simulations and we generated new samples of different sizes to keep them as independent as possible, especially for the convergence study (that has been added to this new version). But your comment is fully relevant and we have added it as a perspective for a catchment-scale application that may be much more computationally costly.

p3 L60-61 "Such approach [...] information on the input.": this sentence if not clear and needs reformulation. Do you mean that this is a GSA method that can be applied 'given data', and does not require a specific structure for the input sample (see e.g. Saltelli et al., 2021)?
Indeed, this sentence meant that the method does not require an input sample with a specific structure nor information about the input distributions. This specific sentence has been removed but the argument in favor of given-data methods now appears in the discussion part so as the reference to Saltelli et al., 2021.

p3 L67: The model used in Vanuytrecht et al. (2014) is not a pesticide model, is it?
Indeed it is not and the reference has been removed.

p3 L69-70 "This qualitative method is based [...] to make clusters appear.": references are missing to support this statement. For instance, Kim et al. (2022, Sect. 4.5) discusses the difficulties in applying Morris method in high dimensions.
Thank you for the valuable reference. However, in the new version of the paper, most of the state-of-the-art on screening, including this sentence, has been removed. Indeed, we have decided to clearly focus the paper on the ranking step so as to provide more results and clearer arguments to choose a specific method.

p11 L239 'accuracy': do you mean precision/robustness (assessed using bootstrapping)? I do not think from your analysis you can infer whether sensitivity indices are accurate. Can we know the 'true' value of the sensitivity indices?

Indeed, we refered to robustness and precision when using the term 'accuracy'. However, I agree that the analysis provided in the initial version did not bring sufficient information to conclude about robustness nor precision of the indices. The new version of the paper now includes results on convergence of the calculated sensitivity indices and the associated error bounds so as to get clear conclusions on these aspects.

p11 L254 "accurately": I would remove this term which I think may be misleading Sobol' indices are typically calculated numerically and not analytically. Can we be sure that the numerical procedure produces accurate sensitivity indices estimates?
Yes, removed as suggested.

p19 L454-455 'It may indicate [...] output variance.': This sentence needs reformulation.
This sentence does not appear anymore in the new version.

p24 L548: I think this sentence needs clarification. The term 'relevancy' is vague. The term 'confidence' is a strong word and I do not think it is appropriate here, given the limitation of these methods discussed in the results section.
The discussion on the methodology has been deeply reviewed and more precise arguments about 'relevancy' and 'confidence' on the methods have been proposed in this section.

p25 L551-552: To put this into context, I suggest to refer to Saltelli et al. (2021), who highlights the benefit of 'given data' sensitivity analysis.
Yes, added as suggested.

2 Reviewer 3

Dear Reviewer 3,

Thank you very much for the careful review and edits to the initital submission. All your comments and questions have been copied hereafter in bold then answered. the revised manuscript is provided as a complement to these answers.

Overall evaluation: The submitted manuscript entitled "How to perform global sensitivity analysis of a catchment-scale, distributed pesticide transfer model? Application to the PESHMELBA model." by Rouzies et al. applies three GSA methods to evaluate the sensitivity of the distributed process-based model to its parameters. The writing is clear and precise, and all sections are understandable. Considering the importance of such analysis for complex hydrologic models, I think the motivation and benefits of this study will be of interest to Geoscientific Model Development readers. Particularly, I like the fact that various GSA methods have been compared in this paper. That being said, the manuscript suffers from some major shortcomings with respect to its novelty and rigor. Here, I outline my comments and suggestions that should allow authors to improve their paper:

Comment 1. The major shortcoming of the paper is that the overall value of this contribution to the hydrologic modelling community is not adequately discussed. The

main contribution of this study is applying three GSA techniques to investigate the role of various parameters in pesticide transfer model. However, its merit over previous attempts is still somehow limited/not well presented. As mentioned by the authors, there are several studies where GSA approach has been applied to explore the factor importance in this context. I am not sure if this and similar studies would add much useful information to the existing body of knowledge on uncertainty analysis, parameter estimation, identifiability analysis, etc. I strongly suggest authors to clearly explain the extend to which this study is adding to the previously presented knowledge in the field (e.g., through new approaches to solving existing problems? etc.).

Having discussed the issue from that point of view, I would rather look at it from another perspective as well. Based on the reported results (Figure 7), overall, the estimated sensitivity indices by RF, HSIC, and Sobol methods are quite different. But, it is not convincing from the paper why one should use HSIC instead of Sobol or RF method. The manuscript correctly mentions the conceptual differences between three methods. For example, HSIC assesses the strength of dependencies between inputs and the output, while Sobol method attributes the variance of the output to variations in inputs or sets of inputs. However, it has not been discussed how this can help modelers/hydrologists with respect to hydrological processes' understanding or model building. To address this issue, I strongly suggest authors provide their "objectives" and "research questions" in the introduction section by bullet points. This can properly highlight the novelty and significance of the study. Furthermore, considering the numerical results, authors should explicitly explain why and how each GSA method might be useful in the context of spatialized pesticide transfer modelling.

As suggested, the objectives and the research questions of the paper have been precised as follows. The main focus of this study is to perform sensitivity analysis of the pesticide transfer model PESHMELBA in order to investigate the role of various parameters in the model. In the previous studies reported in this field, the Sobol method is commonly used for ranking. However, in such studies the number of parameters for ranking is limited (<25) while the number of simulations available is quite high (>10,000). In the case of PESHMELBA model and of similar distributed, multi-processes models, the structure and the spatialized aspect of the model imply a high number of input parameters ($\approx$ 150) and a very limited number of available simulations (<5,000) due to their high computational cost. These constraints imply that classical approaches for GSA cannot be applied as it is to the PESHMELBA model. The objective of this study is then to identify an adequate approach that suits PESHMELBA constraints to perform global sensitivity analysis. The novelty of this study is that we test several low computational budget methods that have been very little used before to perform GSA of pesticide transfer models. The relevancy of these new methods is assessed regarding the following aspects :

- which information do they bring about sensitivity and more generally about physical processes envolved in pesticide transfer and transformation ?
- how robust are they in the case of small sample size ?

The introduction has been deeply reviewed to highlight the novelty of the study and the objectives and research questions as formulated above.
In addition, the discussion section has also been reviewed in order to draw clear conclusions about why and how each method might be useful. Having clarified the objectives of the study, the HSIC measure is particularly examined regarding the criteria above. As you suggested, its interest to help modelers with respect to hydrological processes is now properly discussed and we particularly

highlight the fact that its lack of interpretability is a significant drawback of such a GSA approach.

Comment 2. In my opinion, another major shortcoming of this paper is that there is no information about the convergence behavior of the GSA algorithms. As authors know, robust sensitivity analysis of the models typically requires many model runs, and hence considerable computational resources. So, due to the high number of model evaluations required by existing sensitivity analysis techniques and the computationally expensive nature of the models, analysts usually tend to conduct sensitivity analysis without evaluating its stability and convergence (for more discussion see, e.g., Sarrazin et al., 2016; Sheikholeslami et al., 2020). It is therefore common to choose the sample size only based on the available computational budget, which in turn can result in lack of robustness, and consequently contaminate the assessment of the sensitivities. In fact, since 5 10 years ago a surge of papers flooded the environmental modelling journals introducing/applying a sensitivity analysis technique to a model without analyzing the robustness and converges of the results. Authors should properly monitor/analyze the convergence properties of the utilized GSA techniques in identifying influential factors, for example by progressively increasing the sample size.
I definitely agree with you. As suggested, results about convergence properties of the tested methods have been added (Section 3.2.3) by progressively increasing the sample size up to the maximum possible sample size compatible with our computational budget (2,000 points for each variable for ranking). As already mentionned in the response to comment 1, convergence properties have been highlighted as a criterion for choosing the most suitable method for PESHMELBA.

Comment 3. There is another important cost-effective strategy in the literature to accelerate GSA of the computationally expensive models, namely given-data approach to GSA (otherwise known as data-driven methods). To improve the literature review and strengthen the discussion part, authors can mention given-data approach in the revised manuscript. For a general review and discussion on these techniques see Sheikholeslami et al. (2021).
If I understood correctly the definition of data-driven methods provided in Sheikholeslami et al. (2021), all the methods we applied in this study are data-driven methods as they do not require a specific sample design. Such an argument in favor of these methods has been added in the discussion part.

Comment 4. Going back to comment 3, I think an insufficient state-of-the-art has been performed in this study. There are many studies that have been previously undertaken to develop efficient screening techniques. Authors should consider existing literature in this context and perform a critical review. One notable example is the grouping approach introduced by Sheikholeslami et al. (2019). This approach uses agglomerative hierarchical clustering to categorize the parameters into distinct groups based on similarities between their sensitivity indices, and then ranks parameters according to importance group e.g., these could be labeled as "strongly influential", "influential", "moderately influential", "weakly influential", and "non-influential") rather than individually (see Huo et al., 2019; Sheikholeslami et al., 2021 for further application of the grouping-based importance ranking approach). Other studies include Tang et al., 2007; Nossent et al., 2011; Touzani and Busby, 2014; Becker et al., 2018; etc.
In this study, screening was only performed as a preliminary step to get a reasonable number of input parameters for ranking. We intentionally kept the part of screening short because the focus of

the paper is rather on ranking. To avoid confusion, we have further shortened this part to keep the paper reasonably short. However, the discussion section now mentions that further research should specifically target the screening exercise, including by exploring the approaches you cited.

Comment 5. While "parameter uncertainty" has been thoroughly analyzed in the paper, I could not find proper description about other important sources of uncertainty, particularly input data uncertainties, e.g., soil type, rain and PET forcing data. I recommend authors to add discussion regarding this important source of uncertainty which can significantly affect the model output variability. In fact, these forcings are typically the outputs of a long and complex modelling chains. Thus, PESHMELBA may simultaneously suffer from model parameter, model structure, and input uncertainties or other systematic uncertainties in precipitation bias correction, the estimation of potential evapotranspiration, or the uncertainty of deriving spatial basin scale meteorological input data.
Yes, added as suggested. We are aware that parameter uncertainty is not the only source of uncertainty that can affect the model but it is the only one that is considered in this study for sake of simplicity. Discussion about other sources of uncertainty such as data and model structure have been added as a perspective in the conclusion.

Comment 6. Most of the existing literature on sensitivity analysis has typically been under the assumption that the controlling factors such as model parameters (processes) are independent, whereas, in many cases, they are correlated, and their joint distribution follows a variety of forms. However, very few studies in the field of water and environmental modeling address this issue. By way of example, Strobl et al. (2007) reported that when using permutation-based mean decrease in prediction accuracy as an importance measure, there might be bias in estimating importance of the correlated variables. Authors should highlight this in the revised manuscript by adding discussion on the significance of correlation effects in the utilized methods and then perhaps propose strategies (in future studies) for properly accounting for correlations in parameter (process) space.
This is a very interesting and relevant remark and we added as a perspective in the conclusion. To keep it synthetic, we mainly mentioned the fact that the indices we used in this study may be meaningless in the case of dependent inputs. The main strategy we propose to deal with dependent inputs consists in using the Shapley effects as this lead is very promising and can be adapted to any type of sensitivity indice (see Da Veiga et al. 2021 for further details).

Comment 7. I could not find any information on training and tuning of RF model. The possible inconsistency in SA results might be due to the issues in fitting RF to the input-output data. I strongly suggest authors provide details of building RF model. Furthermore, it's not clear if RF was fitted on scalar quantities or temporal series. Without this information, results are not reliable and cannot be validated.
Details on RF building are provided l.349 and l.350 : The randomForestSRC R package (Ishwaran and Kogalur, 2020) was used to train RF and the number of trees used for training was set to 500. As suggested, we also clarified the fact that the variables used for training were all scalar quantities and that one RF was trained for each variable on each plot.

Comment 8. It would be interesting to see results of parameter ranking as well. Although these methods estimate different values for sensitivity indices in some cases,

the ranking provided by these methods may be much more similar. Note that, in complex models, when the number of parameters is very large, we are typically not interested in an exact values of sensitivity indices. Instead, it may be more profitable to use the available computational budget to rank parameters in order of importance, e.g., "strongly influential", "moderately influential", and "non-influential".
In this case, we are interested in providing a quantified ranking rather than an agglomerative hierarchical clustering because a side objective is to focus parameter characterization efforts. As suggested we have added some results about parameter ranking together with comparison of rankings (Figure 6) and convergence monitoring (Figures 7 and 8).

Comment 9. A possible direction for future research is to evaluate how sensitivity analysis results change by changing the selected parameter distributions (normal, log-normal, uniform,...) since there is an unavoidable uncertainty associated with defining feasible ranges of parameters.
It is indeed a possible direction for future research but we chose to prioritize the question of dealing with dependent inputs in the conclusion part as it has turned a burning question in the community of sensitivity analysis during the last years.